Hole Injection SiO₂ Breakdown Model for Very Low Voltage Lifetime Extrapolation

Klaus F. Schuegraf and Chenming Hu

Abstract—In this paper, we present a model for silicon dioxide breakdown characteristics valid for a thickness range between 25 Å and 130 Å, which provides a method for predicting dielectric lifetime for reduced power supply voltages and aggressively scaled oxide thicknesses. This model, based on hole injection from the anode, accurately predicts QBD and fBD behavior including a fluence in excess of $10^7$ C/cm² at an oxide voltage of 2.4 V for a 25 Å oxide. Moreover, this model is a refinement of and fully complementary with the well known $1/E$ model, while offering the ability to predict oxide reliability for low voltages.

I. INTRODUCTION

FUTURE generation integrated circuits will operate in the presence of reduced supply voltages, thereby providing reduced power consumption as well as assuring adequate circuit reliability in the presence of ever increasing circuit densities. Aggressive scaling of oxide thickness is necessary for enhancing circuit speed, thereby offsetting the disadvantage of supply voltage scaling. Electric fields in the gate insulator of scaled devices are expected to rise with scaling [1] and therefore pose significant reliability concerns. Time-dependent dielectric breakdown, in particular, is a major cause of circuit failure [2]. A $1/E$ model has been effective in formulating a procedure for the reliability assessment of dielectrics in an operating environment where the supply voltage is 5 V [3]. However, to meet future needs of dielectric reliability assurance, a good physical understanding of thin and ultra-thin oxide breakdown at low voltages, 3.3 V and below, is required.

This paper presents a model for the intrinsic low voltage breakdown characteristics, both charge to breakdown, $Q_{BD}$, and breakdown lifetime, $f_{BD}$, of silicon dioxide. The term intrinsic breakdown refers to the breakdown of an oxide that is free of defects, thereby reflecting the practical upper limit of oxide reliability. In practice, intrinsic breakdown is studied by examining oxides having small enough areas such that the breakdown reliability is independent of the area. The intrinsic breakdown reported here corresponds to less than 400 μm² capacitor and transistor test structures. Previous research [4]–[7] has demonstrated the interaction of holes with the oxide in initiating the damage which leads to catastrophic oxide failure. This paper proposes a quantitative model for the physical process responsible for hole generation and injection into silicon dioxide during high field electrical stress. The strength of this model lies in its ability to predict oxide breakdown at the very low operating voltages which are being introduced in state-of-the-art VLSI circuits. This paper also models the enhanced low voltage leakage current due to direct tunneling in ultra-thin oxides.

II. EXPERIMENT

The devices used in this study were either capacitors or transistors with gate oxide thickness varying between 25 Å and 130 Å. Oxidation process was dry oxidation at 750–950°C. All devices have an in situ phosphorus doped polysilicon gate deposited at 605°C. The capacitors were on n-type (100) 8–12 Ω-cm substrate. Constant-voltage stressing of capacitors was performed using a HP4140B picoammeter. Constant-voltage hole separation experiments [7], [8] with transistors were performed using a HP4145B parameter analyzer. All the instruments were controlled by a PC using GPIB interface.

III. ANODE HOLE INJECTION MODEL

We propose that breakdown is a two stage process [3], divided between the time where the oxide is slowly damaged under electrical stress and a much shorter, rapid runaway process, where a rapid final acceleration of damage due to electrical and/or thermal runaway leads to the formation of a permanent conductive path through the oxide. Since the first process dictates the breakdown time, our model addresses the period of accumulation of damage. Recent studies [4], [7] clearly show that hole transport through the oxide precedes breakdown, indicating that holes cause damage to the oxide. This conclusion is not at all surprising when one realizes that holes have been found to generate bulk oxide electron traps [9]–[12] and interface traps [13]. Therefore all models of oxide breakdown variably emphasizing the roles of electron trapping [14], interface trapping [13], hole trapping [3] or resonant tunneling through electron traps [15] can all be reconciled by recognizing the role of hole injection. This model addresses the generation of holes for an oxide biased into the high field tunneling regime as shown in Fig. 1, representing the conditions of accelerated test for determining oxide reliability.

A fraction of the tunneling electrons reaching the anode are able to elastically transfer their entire energy to a deep valenceband electron [16]–[19]. Such an electron is promoted to the lowest available electron energy state, that is, the conduction
band edge of the anode, thereby creating a “hot” hole, which tunnels back into the oxide. These injected holes act to increase the current density (at localized spots), probably through hole-induced trap generation [9]–[12] (see Section VII), until the final runaway process leads to catastrophic breakdown.

Mathematically, the hole tunneling current is given as

\[ J_h = \alpha_p J_n \Theta_p, \]

where \( J_n \) is the incident electron tunneling current, \( \alpha_p \) the probability for a hole to be generated and to tunnel through the barrier. The quantum efficiency of the hole generation process is

\[ J_p = \alpha_p \Theta_p = \alpha_p \exp \left( \frac{-B}{E_{ox}} [\Phi_p(V_{ox})]^\frac{3}{2} \right) \]  

(1)

where \( B = 8\pi \sqrt{2 m_p \Phi_h / 3 h q \Phi_p} \), \( m_p \Phi_h \) is 0.2 times \( m_0 \) [8] and

\[ q \Phi_p = E_{ox} \Phi_p \sin \Phi_0 - q \Phi_0 \]  

(2)

The energy gained from the oxide field before arrival at the anode in the Fowler–Nordheim tunneling regime, where \( V_{ox} > \Phi_p \), calculated by means of a phenomenological energy relaxation model [20], is given as

\[ E_{gain} = \Phi_0 + (E_{ox} \lambda) \left[ 1 - \exp \left( -\frac{1}{\lambda} \left( V_{ox} - \Phi_0 / E_{ox} \right) \right) \right] \]  

(3)

where \( \lambda = 15 \text{Å} \) is the mean free electron scattering length in the oxide conduction band [21]. In contrast, in direct tunneling, i.e., for \( V_{ox} < \Phi_0 \), electrons do not experience such scattering, thus the arrival energy of electrons at the anode is simply

\[ E_{gain} = V_{ox} \text{ for } V_{ox} < \Phi_0. \]  

(4)

Using NMOS transistors biased as in Fig. 2, the tunneling electron current can be easily measured as gate current and the anode-injected hole current can be measured as the substrate current [7], [8]. Fig. 3 shows that the tunnel current density \( J_n \) decreases by over an order of magnitude during the oxide lifetime due to electron trapping. Moreover, as more clearly illustrated in Fig. 4, the quantum hole generation efficiency remains constant for the duration of the oxide lifetime showing that the hole generation rate is strictly determined by the applied bias, consistent with (1) for \( \alpha_p \approx 0.08 \).

The injected hole quantity increases with time as \( Q_p(t) = \int_0^t J_p(t) dt \) until a critical hole fluence, \( Q_p \), is reached, marking the breakdown event. The charge to breakdown, \( Q_{BD} \), follows as

\[ Q_{BD} = \frac{Q_p}{\alpha_p} \left( \frac{B}{E_{ox}} [\Phi_p(V_{ox})]^\frac{3}{2} \right) \]  

(5)

Fig. 5 shows that the critical hole fluence at breakdown, \( Q_p \), is independent of the stress voltage, while \( Q_{BD} \) decreases with increasing stress voltage according to (5). Fig. 6 shows the excellent ability of (5) to predict intrinsic charge to breakdown for oxide thickness varying between 25 and 100 Å. The low voltage predictive ability of (5) attests to the hypothesis of hole injection since tunneling electrons cannot gain the energy necessary for impact ionization in the silicon dioxide at these low voltages. The rapidly rising \( Q_{BD} \) behavior in thinner oxides can be attributed to the fact that the hot hole energy, \( E_{gain} \), becomes more sensitive to \( V_{ox} \) when scattering becomes weaker. The formalism leading to (5) may also be used in modeling thin oxide \( Q_{BD} \) derived from constant current.
Fig. 5. Charge to breakdown, \( Q_{BD} \), and hole fluence to breakdown, \( Q_p \), for a 82 Å oxide, showing that \( Q_{BD} \) increases with decreasing stress voltage, while \( Q_p \) remains constant.

![Graph showing oxide voltage and charge](image)

Fig. 6. Voltage dependence of thin oxide charge to breakdown, demonstrating ability to extrapolate to low voltages.

![Graph showing inverse field and breakdown time](image)

Fig. 7. Comparison of experimentally measured \( Q_p \) values using hole separation and those deduced from fitting capacitor breakdown data.

![Graph showing hole charge to breakdown vs oxide thickness](image)

Fig. 8. I-V characteristics for 51 Å oxide capacitors with in situ doped n+ polysilicon gate. The corrected data for substrate emission and gate emission is computed according to the rules in Table I.

![Graph showing current density vs voltage](image)

Next, quantitative models for Fowler-Nordheim and direct tunneling are introduced.

A. Polysilicon Depletion

In order to model the leakage currents of thin oxides, it is necessary to relate the oxide voltage to the applied bias. Fig. 8 compares the measured positive bias (substrate electron emission: SE) and negative bias (gate electron emission: GE) tunneling currents for a thin oxide capacitor. Whereas thicker oxides display parallel tunneling characteristics with an offset due to band-bending [22], this data shows that the measured J-V characteristics in very thin oxides are not parallel, which can be attributed to the depletion of the heavily in-situ doped n+ gate polysilicon [23]–[25]. Computation of the oxide voltage for each curve by subtracting for band bending in the poly-gate and the substrate reduces them to a single tunneling characteristic dependent only on \( V_{ox} \), consistent with the notion that tunneling should be a unique function of the oxide voltage and thickness, independent of polarity. The polysilicon band bending induces a voltage drop correcting the GE and SE curves under the transformations given in Table I [22], [25] where \( V_{poly} = \frac{q^{2}}{2q^{2}N_{D}} \) until \( V_{poly} \) is pinned at 1.12 V due to the saturation of band bending in strong inversion. The maximum electrically active doping concentration appears to be bounded by approximately \( 5 \times 10^{19} \) cm\(^{-3} \) for this in-situ doped polysilicon, with even lower dopant activation for implanted polysilicon [23], [25]. The single tunneling characteristic dependent only on \( V_{ox} \) attests to the applicability of the depletion approximation [24]. Fig. 9 shows that (5) is valid for predicting \( Q_{BD} \) for either positive or negative bias, providing the oxide voltage is calculated consistently with Table I. However, \( Q_p \) for the gate emission case (negative gate voltage) was found to be 2–5 times lower than its substrate emission (positive voltage) counterpart. This indicates that the poly-SiO\(_2\) interface is not as robust as the Si-SiO\(_2\) interface.

B. Thin Oxide Conduction Model

This section introduces a model for quantitative modeling of the tunnel current, \( J_n \), of Fig. 8. Fig. 10 illustrates the difference between direct and Fowler–Nordheim (FN) tunneling.
Fig. 9. Shows that Anode Hole Injection Model accurately predicts breakdown for positive and negative bias stresses.

<table>
<thead>
<tr>
<th>$V_{OX}$</th>
<th>$V_{G}$</th>
<th>$V_{poly}$</th>
<th>$V_{poly}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$+V_{G}$</td>
<td>n+ poly on p sub</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>$-V_{G}$</td>
<td>n+ poly on n sub</td>
<td>1.2</td>
<td>1.2</td>
</tr>
</tbody>
</table>

The standard FN expression (26) is

$$J_n = AE_{ox}^2 e^{-\frac{\Phi_b}{V_{ox}}}$$

where $B = \frac{8\pi^2 n_m^2 \Phi_{ox}^{3/2}}{3h}$, and the electron effective mass, $m_{ox}$, is 0.5 $m_e$. Equation (6) represents tunneling through the triangular potential barrier of Fig. 10(a), valid for $V_{ox} > \Phi_b$. However, for $V_{ox} < \Phi_b$, where the tunneling barrier is trapezoidal as in Fig. 10(b), (6) is no longer valid but becomes [27],

$$J_n = AE_{ox}^2 \left(\frac{\Phi_b}{V_{ox}}\right) \left(2\frac{\Phi_b}{V_{ox}} - 1\right) e^{-\frac{n \left[ 1 - \left(1 - \frac{V_{ox}}{\Phi_b}\right)^{3/2}}{V_{ox}}}.}$$

This analytical formula does not approach zero as $V_{ox}$ approaches zero. Therefore, it does not apply to the regime of $V_{ox} < 1$ V. An alternative approximation that overcomes this limitation is

$$J_n = AE_{ox}^2 e^{-\frac{n \left[ 1 - \left(1 - \frac{V_{ox}}{\Phi_b}\right)^{3/2}}{V_{ox}}}.}$$

Fig. 11 demonstrates the utility of this model, highlighting the dramatic effect the changed barrier shape, from triangular to trapezoidal, exerts on increasing the leakage current for $V_{ox} < \Phi_b$. Observation of direct tunneling is generally limited to oxides thinner than 50 Å because the tunneling probability for thicker oxides is small. Thus experimental constraints (i.e., current sensitivity of measurement equipment and capacitor area) limit observation of direct tunneling currents.

V. BREAKDOWN LIFETIME

Combining this closed form current model of (6) and (7) with the $Q_{BD}$ model expression of (5) leads to a simple model to predict oxide lifetime, i.e., $t_{BD}=Q_{BD}/J_n$. Although carrier trapping can lead $J_n$ to increase or decrease with stress time, this trapping is negligible for thin oxides so that $t_{BD} \approx Q_{BD}/J_n$. Fig. 12 demonstrates exceptional agreement between $t_{BD}$ theory and intrinsic breakdown data for thin oxide samples, highlighting the ability of the anode hole injection model to predict oxide lifetime at low operating voltages. Whereas the hole-induced $1/E$ extrapolation model predicts an constant extrapolation slope of 350 MV/cm [3], [28], Fig. 13 illustrates that log($t_{BD}$) may still be linearly extrapolated with respect to $1/E$, albeit with increasing extrapolation slope for decreasing oxide thickness. Fig. 14 shows this lifetime extrapolation slope follows

$$G = 400 + 190 \exp \left( -\frac{X_{ox} - 39}{\lambda} \right) \left[ \text{MV/cm} \right]$$

for oxides thickness greater than 39 Å, with $\lambda = 15$ Å. For thinner oxides, the extrapolation slope is about 650 MV/cm. Despite this increase in extrapolation slope, the maximum acceptable oxide field for 10 year lifetime varies by only about 10% for a wide range of oxide thickness as shown in Fig. 13.
Fig. 12. Voltage dependence of breakdown lifetime for thin oxides predicted down to low voltages using Anode Hole Injection Model.

Fig. 13. Inverse field dependence of breakdown lifetime is linear with increasing extrapolation slope for thinner oxides.

Fig. 14. Thickness dependence of inverse field oxide breakdown lifetime slope. Slope increases for thinner oxides due to reduced hole generation efficiency.

VI. SUPPLY VOLTAGE LIMITS

Fig. 15 examines the role of two competing criteria, intrinsic TDDB (10 Year Lifetime at 25°C) and leakage current (0.1 pA/µm²), in determining maximum acceptable operating voltage. Although these two criteria track each other in scaling from 130 Å to 60 Å, further scaling leads these criteria to diverge. The enhanced leakage current due to direct tunneling may require a rapid derating of supply voltage, posing an additional constraint on aggressively scaled oxide technologies below 40 Å. Different leakage requirements may relax the constraint on minimum oxide thickness; however, the rapid increase in direct tunneling current below 40 Å requires rapid supply voltage derating regardless of how much leakage current is permissible. While intrinsic breakdown sets the ultimate reliability limit, in practice, certain thickness margin is needed to cushion against defects whose breakdown is modeled by an effectively thinnest spot [29], [30]. In practice, the ability to manufacture high quality, low defect ultrathin oxides will limit the usable oxide thickness for a given operating voltage.

VII. DISCUSSION

While the preceding discussion focused on the quantitative modeling of thin oxide breakdown characteristics as a function of voltage and thickness, it only hints at the physical mechanism responsible for SiO₂ breakdown. During the course of electrical stress, oxide damage manifests itself as generated interface traps [13], generated bulk traps [9]–[12], [14], low level stress induced leakage [31], [32], and finally, catastrophic breakdown.

Many studies have correlated a particular observation of these manifestations to oxide breakdown and drew conclusions about the cause of oxide wearout. For instance, Harari noted a fixed amount of electron trapping before breakdown and reasoned that electron trapping establishes high internal fields which lead to material rupture. The correlation between electron trapping and breakdown fails to account for the role of holes in the breakdown process. Hot-hole injection by non-Fowler–Nordheim processes [4]–[6], [33] show that hole fluence, not electron fluence, is correlated to oxide breakdown. Hole fluence is also much more deleterious to oxide integrity than electron fluence in that the oxide can sustain a fluence of about 100 times more electrons than holes. Others [34], [35] have suggested that current density can be enhanced by hole trapping and this process leads to oxide wearout. This scenario fails to account for the fact that trapped holes are converted into electron traps [9], [36] and that hole trapping eventually leads to net negative charge trapping in the oxide.

Others have shown the correlation between interface state generation rate and breakdown under static Fowler-Nordheim stress [13], [37], that is, the higher the Nᵢ generation rate, the lower the charge to breakdown. Yet, dynamic bipolar stress generates considerably more interface states than static stress, while also yielding a longer, not shorter, breakdown lifetime [38], [39]. Moreover, Chen [4] showed that large interface
state generation under channel hot-carrier stress did not degrade oxide breakdown integrity. These experiments serve to decouple the correlation between interface state generation and breakdown.

Breakdown appears to be a localized process triggered by a local increase in stress current density. Although the small area of the locally large current density precludes direct measurement of current enhancement, the study of low level stress induced leakage [31, 32, 40] seems to indicate a conduction mechanism of trap assisted tunneling through neutral traps. Such neutral traps [9] created by a two step process of hole trapping and subsequent recombination with electrons [10] may serve to explain the localized increase in current density, be it through the mechanisms of charge assisted tunneling [34], trap assisted tunneling [41] or a closely related resonant tunneling model of breakdown [15]. Since hole fluence leads to the generation of bulk and interface electron traps, it is not surprising that there is often, though not always, a correlation between oxide breakdown and electron trapping or interface trap generation.

Others believe that heating is the final cause of oxide breakdown. A post-breakdown material study reveals that breakdown is a point of localized melting [42], reflecting a large energy discharge at breakdown. Another model proposes that bond breaking by the energetic tunneling electrons forms a conductive path from anode to cathode [43]. Breakdown is the discharge of the capacitor charge through this conductive channel. However, since the discharge process is virtually instantaneous, compared with the damage accumulation time, the length of the discharge event should not determine the breakdown lifetime.

The anode hole injection model includes and integrates all these diverse and valid observations about the breakdown process. Breakdown is a two-step process. The first part is a time, possibly spanning years, where the oxide is slowly damaged under electrical stress. The second step is a very short runaway event, on the order of perhaps microseconds, where a rapid final acceleration of damage due to electrical and/or thermal run-away leads to the formation of a permanent conductive path through the oxide. The injection of holes into the oxide is a precursor to oxide breakdown. Hole injection and trapping lead to the generation of both bulk electron and interface traps [9–12, 36]. Therefore, hole injection, electron trapping, and interface trapping can all be correlated to oxide wearout unless special situations such as AC stress are considered. The neutral traps, so produced, can enhance current density through the oxide. The enhanced current density generates more damage. This positive feedback process causes breakdown to occur at a weak spot.

VIII. CONCLUSION

We have presented a new quantitative model for silicon dioxide breakdown, based on the concept of anode hole injection, valid for predicting insulator reliability performance for thicknesses between 25 and 130 Å. This model is suitable for predicting dielectric lifetime for reduced supply voltages and aggressively scaled oxide thicknesses. Additionally, a quantitative model for enhanced low voltage silicon dioxide leakage current due to the direct tunneling mechanism shows this current may pose additional oxide scaling constraints for oxides thinner than 40 Å. This criterion along with the manufacturability of defect-free ultra-thin silicon dioxide will ultimately determine the usable oxide thickness and supply voltage—and hence the CMOS speed performance.

REFERENCES


Klaus F. Schuegraf received the B.S. degree from Rensselaer Polytechnic Institute, Troy, NY, the M.S. degree from Stanford University, Stanford, CA, and the Ph.D. from the University of California, Berkeley, in 1987, 1988, and 1994, respectively, all in electrical engineering.

From 1987 to 1989 he was a Member of the Technical Staff at Bell Communications Research in Red Bank, NJ. In 1994, he joined the Research and Development Department at Micron Semiconductor, Boise, ID, where he is responsible for high-density DRAM-cell process integration. His research interests include properties and applications of thin dielectrics design for reliability, and process simplification.

Chenming Hu received the B.S. degree from the National Taiwan University and the M.S. and Ph.D. in electrical engineering from the University of California, Berkeley, in 1970 and 1973, respectively.

From 1973 to 1976, he was an Assistant Professor at the Massachusetts Institute of Technology. In 1976 he joined the University of California, Berkeley, as professor of Electrical Engineering and Computer Sciences. He is also Director of the Industrial Liaison Program. While on industrial leave from the University in 1980–1981, he was manager of nonvolatile memory development at National Semiconductor. Since 1973 he has served as a consultant to the electronics industry. He has also been an advisor to many government and educational institutions. His present research interests include VLSI devices, silicon-on-industry devices, hot electron effects, thin dielectrics, electromigration, circuit reliability simulation and nonvolatile semiconductor memories. Dr. Hu has also conducted research on electro-optics, solar cells and power electronics. He has been awarded several patents on semiconductor devices and technology. He has authored or co-authored three books and over 350 research papers. He has delivered more than a score of keynote addresses and invited papers at scientific conferences and has received many best paper awards.

Dr. Hu is an Honorary Professor of Beijing University and Tsinghua University, China, and of the Chinese Academy of Science. Professor Hu is a fellow of the Institute of Electrical and Electronics Engineers. He was Board Chairman of East San Francisco Bay Chinese School from 1988 to 1991. He received the 1991 Design News Excellence in Design Award and the 1991 Semiconductor Research Corporation Technical Excellence Award for leading the development of the IC reliability simulator, BERT.